Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерних наук
(повна назва)

Кафедра програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження моделей прогнозування захворювань для раннього виявлення ризиків та покращення медичної діагностики
(тема)

Виконав:
здобувач 2 року навчання
групи ІПЗм-23-1

Катерина ПОТЬОМКІНА
(Власне ім'я, ПРІЗВИЩЕ)

Спеціальність 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)
Тип програми освітньо-наукова

Керівник проф. Кирило СМЕЛЯКОВ
(посада, Власне ім'я, ПРІЗВИЩЕ)

Допускається до захисту
Зав. кафедри _____ Кирило СМЕЛЯКОВ
(підпис) (Власне ім'я, ПРІЗВИЩЕ)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерних наук
Кафедра програмної інженерії
Рівень вищої освіти другий (магістерський)
Спеціальність 121 – Інженерія програмного забезпечення
Тип програми освітньо-наукова програма
Освітня програма Інженерія програмного забезпечення
(шифр і назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«___» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Потьомкіній Катерині Олексіївні
(прізвище, ім’я, по батькові)

1. Тема роботи «Дослідження моделей прогнозування захворювань для раннього виявлення ризиків та покращення медичної діагностики »
Затверджена наказом по університету від 15.04. 2025р. № 290 Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 20.06.2025
3. Вихідні дані до роботи опис досліджуваних моделей машинного навчання , опис набору даних для тренування та тестування моделей, мова програмування Python, середовища розробки PyCharm Community Edition 2024
4. Перелік питань, що потрібно опрацювати в роботі
аналіз моделей,які використовуються для прогнозування , вибір підходящих даних про медичні показники людей з певною хворобою, написання програмних рішень, проведення експериментів та аналіз отриманих результатів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Отримання завдання	16.04.2025	виконано
2.	Аналіз предметної галузі і постановка задачі	17.04.2025	виконано
3.	Назва за розділами теоретичного і практичного дослідження	20.04.2025	виконано
4.	Назва за розділами теоретичного і практичного дослідження	22.04.2025	виконано
5.	Підготовка до апробації результатів дослідження. Публікація матеріалів	04.05.2025	виконано
6.	Назва за розділами теоретичного і практичного дослідження	10.05.2025	виконано
7.	Підготовка пояснювальної записки	18.05.2025	виконано
8.	Підготовка презентації та доповіді	05.06.2025	виконано
9.	Перевірка на плагіат	14.06.2025	виконано
10.	Нормоконтроль	15.06.2025	виконано
11.	Рецензування	17.06.2025	виконано
12.	Попередній захист	17.06.2025	виконано
13.	Занесення диплома в електронний архів	18.06.2025	виконано
14.	Допуск до захисту у зав. кафедри	18.06.2025	виконано

Дата видачі завдання 16.04.2025р

Студент (ка) _____ Катерина ПОТЬОМКІНА
(підпис)

Керівник роботи _____ проф. Кирило СМЕЛЯКОВ
(підпис) (посада, Власне ім’я, ПРІЗВИЩЕ)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка містить: 61 с., 1 рис., 2 табл., 37 джерел,10 формул

АЛГОРИТМИ, ДЕРЕВА РІШЕНЬ, ЗАХВОРЮВАННЯ, ЛОГІСТИЧНА РЕГРЕСІЯ, МАШИННЕ НАВЧАННЯ, МОДЕЛІ ПРОГНОЗУВАННЯ, XGBOOST

Об’єктом дослідження є моделі машинного навчання.

Метою роботи є проведення дослідження та порівняння моделей машинного навчання, які використовуються для прогнозування захворювань на основі відкритих датасетів.

Методами розробки та проектування є тренування різних моделей відкритими даними про хвороби для проведення дослідження шляхом порівняння результатів використання обраних алгоритмів .

У результаті кваліфікаційної роботи було досліджено та проведено аналіз трьох моделей: Логістична регресія , Дерева рішень та XGBoost.

MACHINE LEARNING, PREDICTION MODELS, DISEASES, LOGISTIC REGRESSION, DECISION TREES, ALGORITHMS, XGBOOST

The subject of this research is machine learning models.
The primary objective of the study is to investigate and compare machine learning models applied to disease prediction using publicly available datasets.
The methodological approach involves the development and training of various prediction models on open-source medical data. The models are evaluated through a comparative analysis of the outcomes produced by selected algorithms.
As a result of this research, three models were examined and analyzed in depth: Logistic Regression, Decision Trees, and XGBoost.

Завідувачу кафедри

ПІ_____

(скорочена назва кафедри)

проф. Кирилу СМЕЛЯКОВУ

(вчене звання, сласне ім’я, прізвище)

ЗАЯВА

щодо самостійності виконання кваліфікаційної роботи та можливості її публікації (та/або публікації анотації кваліфікаційної роботи) в електронному архіві відкритого доступу ElAr KhNURE

Я, Потьомкіна Катерина Олексіївна_____

(прізвище, ім’я, по батькові)

здобувач вищої освіти на другому (магістерському) рівні вищої освітиак адемічної групи ІПЗм-23-4

кафедра_____ програмної інженерії _____,

(повна назва кафедри)

заявляю: моя кваліфікаційна робота на тему

Дослідження моделей прогнозування захворювань для раннього виявлення ризиків та покращення медичної діагностики.

(назва роботи)

що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в репозиторії "ElArKhNURE". Погоджуюся з авторським договором, відповідно до Положення про репозиторій ХНУРЕ "ElArKhNURE". Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлений (а) з вимогами академічної доброчесності, згідно з якими виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

Дата Підпис

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

ADA – Американська діабетична асоціація (American Diabetes Association)

CART – Алгоритм Classification and Regression Trees

GLM – Сімейство узагальнених лінійних моделей (Generalized Linear Models)

IDF – Міжнародна діабетична федерація (International Diabetes Federation)

LIME – Local Interpretable Model-agnostic Explanations

ML – Machine Learning

SHAP – SHapley Additive exPlanations

XGBoost – Extreme Gradient Boosting

ВСТУП

Машинне навчання (ML) стало рушійною силою в сучасній медицині, кардинально змінюючи підходи до прогнозування, діагностики та планування лікування. Зростання обсягів медичних даних у поєднанні з розвитком обчислювальних можливостей створило унікальні можливості для розробки складних моделей прогнозування, які допомагають лікарям ухвалювати точніші та своєчасніші рішення щодо догляду за пацієнтами.

Значення машинного навчання у прогнозуванні захворювань важко переоцінити. Традиційна діагностика в медицині часто базується на симптомах, лабораторних тестах і досвіді лікаря. Однак такий підхід може пропустити тонкі закономірності або ранні ознаки, які є критичними для своєчасного втручання. Алгоритми машинного навчання вирізняються здатністю аналізувати великі обсяги даних, виявляючи взаємозв’язки та фактори ризику, які не завжди очевидні за допомогою традиційних методів.

Особливу увагу привертає застосування ML у діагностиці діабету. Це складний метаболічний розлад, що залежить від багатьох чинників, зокрема генетики, способу життя, екології та супутніх захворювань. За оцінками ВООЗ, понад 422 мільйони людей у світі страждають на діабет, і ця цифра постійно зростає. Зважаючи на серйозні ускладнення, які виникають через пізню діагностику, раннє прогнозування стає вкрай важливим для ефективного лікування та покращення здоров'я пацієнтів.

Моделі машинного навчання для прогнозування діабету враховують різноманітні показники, такі як рівень глюкози в крові, антропометричні дані, спадковість, спосіб життя та демографічну інформацію. Їхня перевага полягає у здатності одночасно обробляти велику кількість змінних і виявляти складні взаємозв’язки, які можуть сигналізувати про підвищений ризик захворювання. Останні дослідження показують, що точність прогнозування таких моделей може перевищувати 85%, що суттєво перевершує традиційні статистичні методи.

Крім діабету, машинне навчання успішно застосовується і для прогнозування інших захворювань, включаючи серцево-судинні хвороби, різні

типи раку, неврологічні розлади та респіраторні захворювання. Алгоритми ML допомагають виявляти закономірності у даних пацієнтів, таких як електрокардіограми, зображення медичних сканувань або поведінкові дані.

Інтеграція ML у медицину має низку переваг, серед яких можливість виявляти фактори ризику на ранніх етапах, більш точна персоналізація лікування, оптимізація використання ресурсів та покращення загальних результатів лікування. Утім, існують і виклики, такі як якість даних, необхідність пояснюваності моделей та потреба в клінічній перевірці. Важливу роль також відіграють етичні аспекти та відповідність регуляторним вимогам.

У перспективі розвиток ML для прогнозування захворювань, особливо таких як діабет, відкриває нові можливості для медицини. Удосконалення методів збору даних і алгоритмів сприятиме створенню більш точних та індивідуалізованих підходів до профілактики та лікування, наближаючи медицину до моделі раннього втручання й ефективного управління здоров’ям.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз предметної галузі дослідження

Цукровий діабет представляє одну з найгостріших медико-соціальних проблем сучасності, характеризуючись епідемічним поширенням та значним тягарем для систем охорони здоров'я у всьому світі. За даними Міжнародної діабетичної федерації (IDF), станом на 2021 рік у світі налічується понад 537 мільйонів дорослих людей із діабетом, що становить приблизно 10,5% від загальної популяції дорослого населення планети. Прогнозні моделі вказують на подальше зростання цього показника до 643 мільйонів осіб до 2030 року та 783 мільйонів до 2045 року, що підкреслює критичність ситуації та необхідність термінових заходів щодо профілактики та раннього виявлення захворювання.[1]

В Україні епідеміологічна ситуація щодо діабету залишається особливо напруженою. Згідно з офіційними статистичними даними Міністерства охорони здоров'я України, станом на кінець 2022 року в країні зареєстровано понад 1,3 мільйона хворих на цукровий діабет, що становить близько 3,2% від загальної чисельності населення. Однак експерти вважають, що реальна кількість хворих може бути значно вищою через високий рівень недіагностованих випадків, який, за оцінками ВООЗ, може сягати 50% від усіх випадків діабету 2 типу.[2] Це означає, що фактична кількість людей із діабетом в Україні може перевищувати 2 мільйони осіб.

Структура захворюваності характеризується переважанням діабету 2 типу, який становить близько 90-95% усіх випадків діабету, тоді як діабет 1 типу складає 5-10%. Особливо тривожною є тенденція до зростання захворюваності серед молодшої популяції, включаючи дітей та підлітків, що пов'язується з погіршенням способу життя, збільшенням поширеності ожиріння та зниженням фізичної активності населення.

Соціально-економічний тягар діабету є надзвичайно високим. Захворювання призводить до розвитку серйозних ускладнень, включаючи діабетичну ретинопатію (яка є провідною причиною сліпоти серед працездатного населення), діабетичну нефропатію (що становить до 40% усіх випадків

термінальної ниркової недостатності), діабетичну нейропатію та макроваскулярні ускладнення у вигляді ішемічної хвороби серця, інсульту та захворювань периферичних судин. Смертність серед хворих на діабет у 2-3 рази вища порівняно із загальною популяцією, а середня тривалість життя скорочується на 8-10 років.

У контексті зростаючого тягаря діабету та необхідності раннього виявлення захворювання особливого значення набувають сучасні технології штучного інтелекту, зокрема методи машинного навчання. Машинне навчання пропонує потужні інструменти для аналізу великих масивів медичних даних, виявлення складних патернів та побудови прогностичних моделей, здатних ідентифікувати осіб з високим ризиком розвитку діабету на доклінічній стадії.[5]

Традиційні підходи до оцінки ризику діабету, такі як шкала FINDRISK або Американська діабетична асоціація (ADA) критерії, базуються на обмеженій кількості факторів ризику та використовують лінійні математичні моделі. Натомість алгоритми машинного навчання здатні одночасно аналізувати сотні параметрів, включаючи демографічні характеристики, антропометричні показники, лабораторні дані, генетичні маркери, особливості способу життя та анамнестичні відомості, виявляючи приховані взаємозв'язки та нелінійні залежності між факторами ризику.

Перевагами застосування машинного навчання в прогнозуванні діабету є висока точність класифікації, здатність до автоматизованої обробки великих обсягів даних, можливість постійного навчання та адаптації моделей до нових даних, а також потенціал для персоналізації ризик-стратифікації з урахуванням індивідуальних особливостей пацієнтів. Сучасні дослідження демонструють, що моделі машинного навчання можуть досягати точності прогнозування діабету на рівні 85-95%, що значно перевищує ефективність традиційних методів.

Крім того, машинне навчання відкриває нові можливості для популяційного скринінгу та стратифікації ризику, дозволяючи медичним закладам ефективно розподіляти обмежені ресурси, зосереджуючи увагу на пацієнтах з найвищим ризиком розвитку захворювання. Це особливо актуально в умовах України, де

система охорони здоров'я стикається з викликами щодо оптимізації ресурсів та підвищення якості профілактичної медицини.[4]

1.2 Постановка задачі

Враховуючи критичну епідеміологічну ситуацію щодо цукрового діабету в Україні та потенціал сучасних технологій машинного навчання для покращення діагностики і прогнозування ендокринних захворювань, актуальною стає задача порівняльного аналізу ефективності різних алгоритмічних підходів для прогнозування ризику розвитку діабету. Метою дослідження є проведення комплексного аналізу трьох базових моделей машинного навчання – логістичної регресії, дерева рішень та XGBoost – для оцінки їх прогностичної здатності, виявлення переваг і недоліків кожного методу та визначення оптимального підходу для клінічного застосування.

Логістична регресія, як один із найпоширеніших методів бінарної класифікації в медичній статистиці, забезпечує інтерпретовані результати через коефіцієнти регресії, що дозволяє клініцистам зрозуміти вплив окремих факторів ризику на ймовірність розвитку діабету. Модель характеризується відносною простотою реалізації, стійкістю до викидів та здатністю працювати з невеликими наборами даних. Однак логістична регресія має обмеження щодо моделювання нелінійних залежностей між змінними та взаємодій між факторами ризику, що може знижувати її точність при аналізі складних метаболічних процесів із багатофакторними взаємозв'язками.

Дерево рішень представляє альтернативний підхід, що автоматично виявляє важливі пороги та взаємодії між змінними, створюючи інтуїтивно зрозумілі правила класифікації. Метод не потребує попередніх припущень щодо розподілу даних та здатний ефективно обробляти як категоріальні, так і числові змінні. Дерева рішень особливо цінні для медичних застосувань завдяки своїй інтерпретованості, дозволяючи лікарям слідувати логіці прийняття рішень щодо ризику діабету. Проте модель схильна до перенавчання, особливо при роботі з

невеликими наборами даних, та може демонструвати нестабільність результатів при незначних змінах у тренувальних даних.

XGBoost (Extreme Gradient Boosting) являє собою сучасний ансамблевий метод, що поєднує множину слабких класифікаторів для створення потужної прогностичної моделі. Алгоритм відомий своєю високою точністю в задачах класифікації та здатністю автоматично обробляти пропущені значення, що є критично важливим для медичних даних. XGBoost ефективно моделює складні нелінійні залежності та взаємодії між метаболічними параметрами, часто демонструючи кращі результати порівняно з традиційними методами. Модель включає регуляризацію для запобігання перенавчанню та забезпечує ранжування важливості ознак. Однак XGBoost характеризується високою обчислювальною складністю, потребує ретельного налаштування гіперпараметрів та може бути менш інтерпретованим порівняно з простішими моделями.

Порівняльний аналіз цих трьох підходів дозволить визначити їх відносну ефективність для прогнозування діабету з урахуванням специфіки медичних даних, включаючи наявність пропущених значень, мультиколінеарність метаболічних показників та необхідність клінічної інтерпретації результатів. Дослідження спрямоване на оцінку точності класифікації, чутливості, специфічності та загальної прогностичної здатності моделей, а також на аналіз їх практичності для впровадження в клінічну практику з урахуванням вимог до інтерпретованості та обчислювальної ефективності, що є критично важливим для розробки ефективних систем підтримки клінічних рішень у сфері ендокринології та діабетології.

2 ОПИС ПРИЙНЯТИХ ПРОЄКТНИХ РІШЕНЬ

2.1 Аналіз моделі Логістична регресія

Логістична регресія представляє собою статистичний метод, який належить до сімейства узагальнених лінійних моделей (GLM) та є одним із найфундаментальніших та широко застосовуваних алгоритмів машинного навчання в медичній статистиці та епідеміології. На відміну від лінійної регресії, яка моделює безпосередній зв'язок між предикторами та цільовою змінною, логістична регресія використовує логістичну функцію (сигмоїд) для моделювання ймовірності належності до певного класу, що робить її ідеальною для задач бінарної та багатокласової класифікації в медичній діагностиці.

Математично логістична регресія базується на логіт-трансформації, яка перетворює ймовірність $p \in [0,1]$ у логарифм відношення шансів (log-odds), що описується формулою (2.1):

$$\text{logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (2.1)$$

де p – ймовірність належності до позитивного класу;

β_0 – константа (intercept);

β_i – коефіцієнти регресії для предикторів;

x_i – значення предикторів;

k – кількість предикторів.

Ймовірність належності до позитивного класу визначається через зворотну логістичну функцію згідно з формулою (2.2):

$$p = \exp(\beta_0 + \Sigma\beta_i x_i) / (1 + \exp(\beta_0 + \Sigma\beta_i x_i)) = 1 / (1 + \exp(-(\beta_0 + \Sigma\beta_i x_i))) \quad (2.2)$$

де exp – експоненціальна функція;

$\Sigma\beta_i x_i$ – сума добутків коефіцієнтів на відповідні предиктори.

Оцінка параметрів здійснюється методом максимальної правдоподібності (Maximum Likelihood Estimation), який максимізує логарифм функції правдоподібності відповідно до формули (2.3):

$$L(\beta) = \Sigma_i[y_i \ln(p_i) + (1-y_i) \ln(1-p_i)] \quad (2.3)$$

де $L(\beta)$ – логарифм функції правдоподібності;

y_i – фактичне значення цільової змінної (0 або 1);

p_i – прогнозована ймовірність для і-го спостереження;

ln – натуральний логарифм;

і – індекс спостереження.

Відношення шансів (Odds Ratio, OR) для кожного предиктора обчислюється за формулою (2.4):

$$OR = \exp(\beta_i) \quad (2.4)$$

де OR – відношення шансів;

β_i – коефіцієнт регресії для і-го предиктора.

Відношення шансів має пряму клінічну інтерпретацію: OR показує, у скільки разів збільшуються шанси настання події при зміні предиктора на одиницю, утримуючи інші змінні постійними

Логістична регресія базується на кількох ключових припущеннях, дотримання яких є критично важливим для отримання валідних результатів у медичних дослідженнях. Першим припущенням є лінійність зв'язку між предикторами та логітом цільової змінної. Це означає, що логарифм відношення шансів повинен лінійно залежати від значень предикторів, що може бути перевірено через аналіз залишків або графічні методи.

Важливим є припущення про відсутність мультиколінеарності між предикторами, оскільки високі кореляції між незалежними змінними можуть призвести до нестабільності оцінок коефіцієнтів та неадекватної інтерпретації

результатів. В медичному контексті це особливо актуально при аналізі лабораторних показників, які часто корелюють між собою через спільні патофізіологічні механізми.

Логістична регресія передбачає незалежність спостережень, що може порушуватися в медичних дослідженнях з кластерним дизайном (наприклад, пацієнти з одного медичного закладу) або повторними вимірюваннями. У таких випадках необхідно використовувати змішані моделі або методи корекції стандартних помилок.

Модель також чутлива до впливових спостережень та викидів, які можуть значно впливати на оцінки коефіцієнтів. В медичних даних такі випадки можуть представляти рідкісні клінічні синдроми або помилки вимірювання, тому їх ідентифікація та аналіз є важливим етапом моделювання.

Логістична регресія має надзвичайно широке застосування в медичній практиці та дослідженнях завдяки своїй здатності моделювати бінарні клінічні виходи та надавати клінічно інтерпретовані результати. В діагностичній медицині модель використовується для розробки правил прогнозування ризику, які інтегрують множину клінічних параметрів для оцінки ймовірності наявності захворювання.

У кардіології логістична регресія широко застосовується для розробки калькуляторів ризику серцево-судинних подій, таких як Framingham Risk Score або SCORE. Ці моделі аналізують традиційні фактори ризику (вік, стать, паління, артеріальний тиск, рівень холестерину) для прогнозування 10-річного ризику інфаркту міокарда або інсульту. Точність таких моделей зазвичай становить 70-80%, що є прийнятним для клінічного застосування.

В онкології логістична регресія використовується для створення прогностичних моделей виживаності, оцінки ризику метастазування та прогнозування відповіді на терапію. Наприклад, модель Oncotype DX для раку молочної залози використовує логістичну регресію для аналізу експресії генів та прогнозування користі від хіміотерапії. Подібні підходи застосовуються для раку простати (Decipher), колоректального раку та інших онкологічних захворювань.

У превентивній медицині та скринінгу логістична регресія допомагає ідентифікувати осіб з високим ризиком розвитку захворювань для цільових втручань. Моделі прогнозування діабету, такі як Finnish Diabetes Risk Score (FINDRISK), використовують логістичну регресію для аналізу демографічних характеристик, антропометричних показників та сімейного анамнезу.

В інфекційних хворобах модель застосовується для прогнозування тяжкості перебігу, ризику ускладнень та смертності. Під час пандемії COVID-19 було розроблено численні логістичні моделі для прогнозування потреби в госпіталізації, ШВЛ та летального виходу на основі клінічних та лабораторних параметрів при надходженні.

Оцінка якості логістичної регресії в медичних застосуваннях включає кілька ключових аспектів: дискримінаційну здатність, калібрацію та клінічну корисність. Дискримінаційна здатність оцінюється через площу під ROC-кривою (AUC-ROC), яка показує здатність моделі розрізняти між позитивними та негативними випадками. Значення AUC > 0.7 вважається прийнятним, > 0.8 - хорошим, а > 0.9 - відмінним для медичних застосувань.

Калібрація характеризує відповідність між прогнозованими ймовірностями та фактичною частотою подій. Вона оцінюється через тест Хосмера-Лемешова, калібраційні графіки та показники типу Brier score. Хороша калібрація критично важлива для клінічного застосування, оскільки лікарі повинні мати можливість довіряти прогнозованим ймовірностям при прийнятті рішень.

Внутрішня валідація здійснюється через методи крос-валідації або bootstrap, які дозволяють оцінити стабільність моделі та уникнути оптимістичних оцінок її ефективності. Зовнішня валідація на незалежних наборах даних є золотим стандартом для підтвердження клінічної застосовності моделі.

Аналіз впливових спостережень включає обчислення статистик типу leverage, Cook's distance та standardized residuals для ідентифікації випадків, які можуть непропорційно впливати на результати моделі. В медичному контексті такі спостереження можуть представляти рідкісні клінічні варіанти або помилки в даних.

Основною перевагою логістичної регресії в медичній практиці є її виняткова інтерпретованість та пряма клінічна релевантність результатів. Коефіцієнти регресії мають чітке статистичне та клінічне значення: $\exp(\beta_i)$ представляє відношення шансів, яке показує, як змінюється ризик події при зміні предиктора на одиницю. Це дозволяє лікарям зрозуміти внесок кожного фактора ризику та приймати обґрунтовані клінічні рішення.

Модель надає прогнозовані ймовірності, які можуть безпосередньо використовуватися для комунікації з пацієнтами та прийняття спільних рішень щодо лікування. Наприклад, лікар може сказати пацієнту: "Ваш ризик серцевого нападу протягом наступних 10 років становить 15%", що є набагато більш інформативним, ніж просте віднесення до групи "високого ризику".

Логістична регресія добре працює з невеликими наборами даних та не потребує великої кількості параметрів для налаштування, що робить її практичною для клінічних досліджень з обмеженими вибірками. Модель стійка до шуму в даних та не схильна до перенавчання при правильному використанні методів регуляризації.

Важливою перевагою є можливість включення взаємодій між змінними через додавання відповідних термів до моделі. Це дозволяє моделювати складні клінічні сценарії, де ефект одного фактора ризику залежить від значення іншого (наприклад, вплив віку може різнитися між чоловіками та жінками).

Обмеження та виклики в медичному застосуванні

Попри численні переваги, логістична регресія має кілька важливих обмежень у медичному контексті. Основним обмеженням є припущення про лінійність зв'язку між предикторами та логітом виходу, що може не виконуватися для складних біологічних процесів. Порушення цього припущення може призвести до неадекватного моделювання та погіршення прогностичної здатності.

Модель може мати труднощі з моделюванням складних нелінійних взаємозв'язків та взаємодій високого порядку між множинними змінними. В медицині такі взаємодії часто існують через складні патофізіологічні механізми, що можуть знижувати ефективність простих логістичних моделей.

Логістична регресія чутлива до мультиколінеарності, що є поширеною проблемою в медичних даних, де різні показники можуть корелювати через спільні біологічні процеси. Високі кореляції між предикторами можуть призвести до нестабільних оцінок коефіцієнтів та труднощів в інтерпретації результатів.

При роботі з незбалансованими класами, що часто зустрічається в медицині при вивченні рідкісних захворювань, стандартна логістична регресія може демонструвати зміщення в бік більшого класу. Це потребує використання спеціальних техніків балансування або коригування порогів класифікації.

Проблема "прокляття розмірності" може виникати при великій кількості предикторів відносно розміру вибірки, що призводить до перенавчання та погіршення узагальнюючої здатності моделі. В медичних дослідженнях це особливо актуально при аналізі геномних даних або при включенні великої кількості потенційних факторів ризику.

2.2 Аналіз моделі Дерево рішень

Дерево рішень представляє собою один із найбільш інтуїтивних та широко застосовуваних методів машинного навчання в медичній діагностиці, що базується на принципі послідовного розбиття множини спостережень на однорідні підмножини за допомогою системи бінарних питань. Алгоритм створює ієрархічну структуру, де кожен внутрішній вузол представляє тест на певну ознаку, кожна гілка відповідає результату тесту, а листкові вузли містять кінцеві рішення щодо класифікації або прогнозу.

Математично процес побудови дерева рішень можна формалізувати як рекурсивний поділ навчальної вибірки D на підмножини D_l та D_r за допомогою функції розбиття $f(x) \leq t$, де x є вектором ознак, а t - пороговим значенням. Оптимальне розбиття вибирається за критерієм максимізації зниження ентропії або індексу Джині згідно з формулою (2.5):

$$\Delta I = I(D) - (|D_l|/|D|)I(D_l) - (|D_r|/|D|)I(D_r) \quad (2.5)$$

де ΔI – зниження міри невизначеності;

I(D) – міра невизначеності множини D;

D_l – ліва підмножина після розбиття;

D_r – права підмножина після розбиття;

|D| – кількість елементів у множині D;

$|D_l|$ – кількість елементів у лівій підмножині;

$|D_r|$ – кількість елементів у правій підмножині.

Для задач класифікації найчастіше використовуються критерії інформаційного виграшу на основі ентропії Шеннона відповідно до формули (2.6):

$$H(S) = -\Sigma_{i=1}^{c} p_i \log_2(p_i) \quad (2.6)$$

де H(S) – ентропія Шеннона для множини S;

c – кількість класів;

p_i – ймовірність класу i в множині S;

$\log_2$ – логарифм за основою 2.6

Альтернативно використовується індекс Джині, що визначається формулою (2.7):

$$G(S) = 1 - \Sigma_{i=1}^{c} p_i^2 \quad (2.7)$$

де G(S) – індекс Джині для множини S;

p_i – ймовірність класу i в множині S;

c – кількість класів.

Процес розбиття продовжується рекурсивно до досягнення критеріїв зупинки, таких як мінімальна кількість спостережень у вузлі, максимальна глибина дерева або відсутність статистично значущого покращення чистоти вузлів.

Існує кілька основних алгоритмів побудови дерев рішень, кожен з яких має специфічні особливості для медичних застосувань. Алгоритм ID3 (Iterative Dichotomiser 3) використовує інформаційний виграш як критерій розбиття та працює виключно з категоріальними ознаками.[19] Його розширення C4.5 підтримує як категоріальні, так і числові ознаки, включає механізми обробки пропущених значень та пост-обрізання для запобігання перенавчанню.[20]

Алгоритм CART (Classification and Regression Trees) є універсальним підходом, що підтримує як задачі класифікації, так і регресії.[21] Для класифікації CART використовує індекс Джині, а для регресії - середньоквадратичну помилку. Особливістю CART є створення виключно бінарних дерев та вбудована процедура обрізання на основі мінімізації помилки крос-валідації.

Сучасні реалізації включають додаткові оптимізації для медичних даних: обробку незбалансованих класів через ваговані функції втрат, підтримку різних типів медичних даних (лабораторні показники, категоріальні діагнози, ординальні шкали тяжкості), інтеграцію з методами імпутації пропущених значень та можливість інкорпорації експертних знань через обмеження на структуру дерева.

Дерева рішень знаходять широке застосування в різних сферах медичної діагностики завдяки своїй природній здатності моделювати діагностичні алгоритми, які використовують лікарі в клінічній практиці. В первинній медичній допомозі дерева рішень ефективно застосовуються для тріажу пацієнтів, дозволяючи автоматизувати процес визначення пріоритетності та маршрутизації хворих на основі симптомів та базових клінічних параметрів.

У кардіології дерева рішень успішно використовуються для діагностики ішемічної хвороби серця, стратифікації ризику серцево-судинних подій та прогнозування ефективності терапевтичних втручань.[22] Модель може інтегрувати результати електрокардіографії, ехокардіографії, стрес-тестів та лабораторних показників для створення комплексного діагностичного алгоритму. Дослідження демонструють точність діагностики на рівні 85-90% для виявлення значущих коронарних стенозів.

В онкології дерева рішень застосовуються для скринінгу злоякісних новоутворень, диференційної діагностики між доброякісними та злоякісними утвореннями, а також для прогнозування відповіді на терапію та віддалених результатів лікування. Особливо ефективним є використання дерев рішень для аналізу гістопатологічних даних, де алгоритм може ідентифікувати ключові морфологічні критерії злоякісності.[23]

У психіатрії та неврології дерева рішень допомагають у діагностиці когнітивних порушень, депресивних розладів та нейродегенеративних захворювань. Модель аналізує результати нейропсихологічного тестування, нейровізуалізації та клінічних шкал для створення алгоритмів диференційної діагностики між різними формами деменції або психічних розладів.[24]

В інфекційних хворобах дерева рішень використовуються для ранньої діагностики сепсису, прогнозування тяжкості перебігу інфекційних захворювань та оптимізації антибіотикотерапії. Алгоритм аналізує клінічні симптоми, лабораторні маркери запалення та результати мікробіологічних досліджень для швидкого прийняття клінічних рішень.[25]

Основною перевагою дерев рішень в медичній діагностиці є їх виняткова інтерпретованість та прозорість процесу прийняття рішень. На відміну від "чорних скриньок" типу нейронних мереж, дерева рішень створюють зрозумілі алгоритми, які лікарі можуть легко слідувати та валідувати на основі свого клінічного досвіду. Це особливо важливо в медицині, де рішення можуть мати життєво важливі наслідки.[26]

Дерева рішень природним чином обробляють як числові, так і категоріальні змінні без необхідності попередньої трансформації даних. Це критично важливо для медичних застосувань, де дані часто включають різнотипні параметри: числові лабораторні показники, категоріальні діагностичні коди, ординальні шкали тяжкості та бінарні індикатори наявності симптомів.

Алгоритм робастний до викидів та не потребує припущень щодо розподілу даних, що робить його придатним для реальних медичних даних, які часто характеризуються неоднорідністю та наявністю атипових значень. Дерева рішень

автоматично виявляють нелінійні взаємозв'язки та взаємодії між змінними, що може розкрити складні патофізіологічні механізми захворювань.

Важливою перевагою є здатність дерев рішень працювати з відносно невеликими наборами даних, що часто зустрічається в медичних дослідженнях, особливо при вивченні рідкісних захворювань. Модель може створювати ефективні правила класифікації навіть за умови обмеженої кількості спостережень.

Попри численні переваги, дерева рішень мають кілька важливих обмежень у медичному контексті. Основною проблемою є схильність до перенавчання, особливо при роботі з зашумленими медичними даними або невеликими вибірками. Перенавчання призводить до створення надмірно складних дерев, які добре класифікують тренувальні дані, але погано узагальнюються на нові випадки.

Дерева рішень можуть демонструвати нестабільність результатів при незначних змінах у тренувальних даних. Це особливо проблематично в медицині, де важливо мати стабільні та відтворювані діагностичні алгоритми. Невелика зміна в наборі даних може призвести до кардинально іншої структури дерева та інших діагностичних правил.

Алгоритм має тенденцію до створення зміщених моделей при роботі з незбалансованими класами, що часто зустрічається в медичних даних, де рідкісні захворювання представлені значно меншою кількістю випадків порівняно зі здоровими особами. Це може призвести до недооцінки ризику розвитку рідкісних, але клінічно значущих станів.

Дерева рішень мають обмежену здатність до моделювання лінійних взаємозв'язків та можуть створювати надмірно складні структури для відносно простих залежностей. Крім того, алгоритм може мати труднощі з обробкою великої кількості ознак, що призводить до фрагментації даних та зниження статистичної потужності.

2.3 Аналіз моделі XGBoost

XGBoost (Extreme Gradient Boosting) представляє собою оптимізовану реалізацію алгоритму градієнтного бустінгу, розроблену Тяньці Ченом у 2016 році, яка стала одним із найефективніших методів машинного навчання для задач класифікації та регресії. Алгоритм базується на концепції ансамблевого навчання, де множина слабких навчальних моделей (зазвичай дерев рішень) послідовно комбінується для створення потужного прогностичного алгоритму. Основна ідея полягає в ітеративному додаванні нових моделей, кожна з яких навчається на помилках попередніх, що дозволяє поступово покращувати загальну точність прогнозування.

Математично XGBoost можна представити як адитивну модель згідно з формулою (2.8):

$$\hat{y}_i = \Sigma_{k=1}^{K} f_k(x_i), \text{ де } f_k \in F \tag{2.8}$$

де $\hat{y}_i$ – прогнозоване значення для і-го спостереження;

K – кількість дерев у моделі;

f_k – k-те дерево рішень;

x_i – вектор ознак для і-го спостереження;

F – простір усіх можливих дерев рішень.

Процес навчання оптимізує регуляризовану цільову функцію відповідно до формули (2.9):

$$L(\varphi) = \Sigma_i\, l(\hat{y}_i, y_i) + \Sigma_k\, \Omega(f_k) \tag{2.9}$$

де $L(\varphi)$ – регуляризована цільова функція;

$l(\hat{y}_i, y_i)$ – диференційована функція втрат;

y_i – фактичне значення для і-го спостереження;

$\Omega(f_k)$ – регуляризуючий терм для k-го дерева;

φ – параметри моделі.

Кожне f_k є окремим деревом з власною структурою листків і ваговими коефіцієнтами, а регуляризуючий терм $\Omega(f_k)$ контролює складність моделі для запобігання перенавчанню.Ключовою особливістю XGBoost є використання градієнтів другого порядку (гессіан) для оптимізації, що забезпечує швидшу конвергенцію та вищу точність порівняно з традиційними методами градієнтного бустінгу. Алгоритм включає розвинену систему регуляризації, яка запобігає перенавчанню через контроль глибини дерев, мінімальної кількості спостережень у листках та L1/L2 регуляризацію вагів.

XGBoost характеризується рядом технічних інновацій, що роблять його особливо придатним для медичних застосувань. Алгоритм реалізує ефективну обробку розріджених даних через column block structure, що дозволяє швидко працювати з великими наборами медичних даних, які часто містять значну кількість пропущених значень. Система автоматично обробляє відсутні дані шляхом вивчення оптимального напрямку розгалуження для кожного вузла дерева.[27]

Особливістю XGBoost є підтримка різних типів цільових функцій, включаючи логістичну регресію для бінарної класифікації (діагностика захворювання/здоровий стан), багатокласову класифікацію (диференційна діагностика між кількома захворюваннями) та регресію (прогнозування числових показників, таких як рівень глюкози або артеріального тиску). Алгоритм підтримує паралельне та розподілене обчислення, що критично важливо для обробки великих медичних баз даних.

Система включає вбудовані механізми контролю перенавчання через early stopping, cross-validation та моніторинг метрик валідації під час навчання. XGBoost забезпечує детальний аналіз важливості ознак через кілька методів: gain (покращення точності від використання ознаки), cover (кількість спостережень, що потрапляють у розгалуження за цією ознакою) та frequency (частота використання ознаки в деревах).

У сфері медичної діагностики XGBoost демонструє винятково високу ефективність завдяки здатності моделювати складні нелінійні взаємозв'язки між

клінічними параметрами, лабораторними показниками та симптомами пацієнтів. Алгоритм особливо цінний для аналізу мультиморбідних станів, де одночасно присутні кілька захворювань з перехресними симптомами та взаємопов'язаними факторами ризику.[28]

В онкології XGBoost успішно застосовується для прогнозування ризику розвитку різних типів раку, включаючи рак молочної залози, легенів, простати та колоректальний рак. Дослідження показують, що алгоритм здатен інтегрувати геномні дані, гістопатологічні зображення, клінічні параметри та епідеміологічні фактори для створення комплексних прогностичних моделей з точністю до 90-95%.[29]

У кардіології XGBoost ефективно прогнозує ризик серцево-судинних подій, включаючи інфаркт міокарда, інсульт та серцеву недостатність. Модель аналізує електрокардіографічні параметри, ехокардіографічні показники, лабораторні маркери та фактори ризику для стратифікації пацієнтів та персоналізації терапевтичних підходів.[30]

В ендокринології алгоритм широко використовується для раннього виявлення цукрового діабету та його ускладнень. XGBoost аналізує глікемічні профілі, антропометричні показники, генетичні маркери та особливості способу життя для ідентифікації осіб з високим ризиком розвитку діабету на доклінічній стадії.[31]

Основними перевагами XGBoost в медичних застосуваннях є висока прогностична точність, що часто перевищує традиційні статистичні методи на 10-20%. Алгоритм ефективно обробляє гетерогенні медичні дані, включаючи числові лабораторні показники, категоріальні діагностичні коди, текстові анамнестичні дані та зображення після відповідної попередньої обробки.

Важливою перевагою є робастність до пропущених даних, що критично важливо в медичній практиці, де неповнота інформації є поширеною проблемою через технічні обмеження, економічні фактори або особливості клінічного обстеження. XGBoost автоматично визначає оптимальну стратегію обробки відсутніх значень без необхідності їх імпутації.

Алгоритм забезпечує ранжування важливості ознак, що дозволяє клініцистам ідентифікувати найбільш значущі предиктори захворювання та оптимізувати діагностичні протоколи. Це особливо цінно для розробки скринінгових програм та протоколів персоналізованої медицини.

Серед обмежень XGBoost слід відзначити відносно низьку інтерпретованість порівняно з простішими моделями, такими як логістична регресія або дерева рішень. Хоча алгоритм надає інформацію про важливість ознак, розуміння конкретних механізмів прийняття рішень залишається складним, що може ускладнювати клінічне застосування в критичних ситуаціях.

Висока обчислювальна складність потребує значних ресурсів для навчання та оптимізації гіперпараметрів, що може бути проблематичним для медичних закладів з обмеженою технічною інфраструктурою. Крім того, модель потребує ретельного налаштування множини гіперпараметрів, включаючи швидкість навчання, глибину дерев, субсемплінг та регуляризацію, що вимагає високої експертизи в галузі машинного навчання.

Сучасні тенденції розвитку XGBoost в медичній діагностиці спрямовані на покращення інтерпретованості через інтеграцію з методами пояснювального штучного інтелекту (Explainable AI), такими як SHAP (SHapley Additive exPlanations) та LIME (Local Interpretable Model-agnostic Explanations). Ці підходи дозволяють отримати детальні пояснення для кожного прогнозу, що є критично важливим для клінічного прийняття рішень та регуляторного схвалення медичних AI-систем.

Перспективним напрямом є розвиток федеративного навчання XGBoost, що дозволить тренувати моделі на розподілених медичних даних без порушення конфіденційності пацієнтів. Це особливо актуально для створення глобальних діагностичних систем, які навчаються на даних з різних медичних установ та регіонів.[32]

Інтеграція XGBoost з методами глибокого навчання відкриває нові можливості для аналізу медичних зображень, електрофізіологічних сигналів та геномних даних.[33] Гібридні архітектури, що поєднують переваги нейронних

мереж для обробки неструктурованих даних та XGBoost для структурованих клінічних параметрів, демонструють перспективні результати в комплексній медичній діагностиці.

Впровадження XGBoost в клінічну практику вимагає розробки спеціалізованих медичних інформаційних систем, що забезпечують інтеграцію з електронними медичними записами, автоматизацію процесів передобробки даних та представлення результатів у зручному для клініцистів форматі.

3 ОПИС ПРОГРАМНОЇ РЕАЛІЗАЦІЇ

3.1 Аналіз набору даних

Для проведення дослідження використовувався набір даних "Diabetes 130-US Hospitals for Years 1999-2008", створений на основі бази даних Health Facts (Cerner Corporation, Канзас-Сіті, Міссурі) — національного сховища даних, яке збирає комплексні клінічні записи з лікарень по всій території Сполучених Штатів. Health Facts є добровільною програмою, запропонованою організаціям, які використовують систему електронних медичних записів Cerner. База даних містить систематично зібрані дані з електронних медичних записів учасницьких інституцій та включає дані про звернення (невідкладна допомога, амбулаторне та стаціонарне лікування), спеціальність лікаря, демографічні показники (вік, стать та раса), діагнози та внутрішньолікарняні процедури, задокументовані кодами МКХ-9-КМ, лабораторні дані, фармацевтичні дані, внутрішньолікарняну смертність та характеристики лікарень.

Набір даних "Diabetes 130-US Hospitals for Years 1999-2008" являє собою витяг з Health Facts, що представляє 10 років (1999-2008) клінічної допомоги в 130 лікарнях та інтегрованих мережах надання медичних послуг по всій території Сполучених Штатів: Середній Захід (18 лікарень), Північний Схід (58), Південь (28) та Захід (16). Більшість лікарень (78) мають ліжковий фонд від 100 до 499 ліжок, 38 лікарень мають менше 100 ліжок, а 14 лікарень мають понад 500 ліжок.

База даних складається з 41 таблиці у схемі факт-вимір та загалом 117 характеристик. База даних включає 74,036,643 унікальних звернень (візитів), що відповідають 17,880,231 унікальним пацієнтам та 2,889,571 лікарям. Оскільки ці дані представляють інтегровані системи охорони здоров'я на додаток до окремих лікарень, дані містять як стаціонарні, так і амбулаторні дані, включаючи відділення невідкладної допомоги, для тієї ж групи пацієнтів.

Створення набору даних відбувалося у два етапи. Спочатку з бази даних було витягнуто цікаві звернення з 55 атрибутами. По-друге, було проведено попередній аналіз та попередню обробку даних, що призвело до збереження лише

тих характеристик (атрибутів) та звернень, які можуть бути використані в подальшому аналізі, тобто містять достатню інформацію.

Інформацію було витягнуто з бази даних для звернень, які задовольняли наступні критерії: це стаціонарне звернення (госпіталізація); це "діабетичне" звернення, тобто таке, під час якого будь-який вид діабету був введений у систему як діагноз; тривалість перебування становила принаймні 1 день і не більше 14 днів; під час звернення проводилися лабораторні дослідження; під час звернення вводилися медикаменти.

101,766 звернень були ідентифіковані як такі, що відповідають усім п'яти критеріям включення та використовувалися для подальшого аналізу. Відбір атрибутів/характеристик проводився клінічними експертами, і залишалися лише атрибути, які потенційно пов'язані з діабетичним станом або лікуванням. З інформації, доступної в базі даних, було витягнуто 55 характеристик, що описують діабетичні звернення, включаючи демографічні дані, діагнози, діабетичні медикаменти, кількість візитів у році, що передував зверненню, та інформацію про платника.

Оскільки дослідження в першу чергу зацікавлене у факторах, що призводять до раннього повторного госпіталізації, атрибут повторного госпіталізації (результат) було визначено як такий, що має два значення: "повторно госпіталізований", якщо пацієнт був повторно госпіталізований протягом 30 днів після виписки, або "інакше", що охоплює як повторну госпіталізацію після 30 днів, так і відсутність повторної госпіталізації взагалі.

Таблиця 3.1 – Таблиця параметрів набору даних (таблиця створена самостійно)

Параметр	Тип	Опис	% відсутніх
Encounter ID	Числовий	Унікальний ідентифікатор звернення	0%
Patient number	Числовий	Унікальний ідентифікатор пацієнта	0%
Race	Номінальний	Значення: кавказець, азіат, афроамериканець, латиноамериканець	2%
Gender	Номінальний	Значення: чоловік, жінка,	0%

Продовження таблиці 3.1

Параметр	Тип	Опис	% відсутніх
		невідомо/недійсно	
Age	Номінальний	Групується за 10-річними інтервалами: [0, 10), [10, 20), …, [90, 100)	0%
Weight	Числовий	Вага в фунтах	97%
Admission type	Номінальний	Ідентифікатор, що відповідає 9 різним значенням (невідкладний, плановий, новонароджений)	0%
Discharge disposition	Номінальний	Ідентифікатор, що відповідає 29 різним значенням (виписаний додому, помер)	0%
Admission source	Номінальний	Ідентифікатор, що відповідає 21 різним значенням (направлення лікаря, швидка допомога)	0%
Time in hospital	Числовий	Кількість днів між госпіталізацією та випискою	0%
Payer code	Номінальний	Ідентифікатор, що відповідає 23 різним значенням (Medicare, самооплата)	52%
Medical specialty	Номінальний	Ідентифікатор спеціальності лікаря (84 різних значення: кардіологія, терапія)	53%
Number of lab procedures	Числовий	Кількість лабораторних тестів, проведених під час звернення	0%
Number of procedures	Числовий	Кількість процедур (окрім лабораторних) під час звернення	0%
Number of medications	Числовий	Кількість різних генеричних назв, введених під час звернення	0%
Number of outpatient visits	Числовий	Кількість амбулаторних vizитів пацієнта в році перед зверненням	0%
Number of emergency visits	Числовый	Кількість візитів швидкої допомоги пацієнта в році перед зверненням	0%
Number of inpatient visits	Числовий	Кількість стаціонарних візитів пацієнта в році перед зверненням	0%
Diagnosis 1	Номінальний	Первинний діагноз (перші три цифри МКХ-9); 848 різних значень	0%
Diagnosis 2	Номінальний	Вторинний діагноз (перші три цифри МКХ-9); 923 різних значення	0%
Diagnosis 3	Номінальний	Додатковий вторинний діагноз (перші три цифри МКХ-9); 954 різних	1%

Кінець таблиці 3.1

Параметр	Тип	Опис	% відсутніх
		значення	
Number of diagnoses	Числовий	Кількість діагнозів, введених у систему	0%
Glucose serum test result	Номінальний	Діапазон результату або якщо тест не проводився (">200," ">300," "normal," "none")	0%
A1c test result	Номінальний	Діапазон результату або якщо тест не проводився (">8", ">7", "normal", "none")	0%
Change of medications	Номінальний	Чи була зміна в діабетичних медикаментах ("change", "no change")	0%
Diabetes medications	Номінальний	Чи призначалися діабетичні медикаменти ("yes", "no")	0%

3.2 Попередня обробка даних

Оригінальна база даних містить неповну, надлишкову та зашумлену інформацію, як і очікується в будь-яких реальних даних. Було виявлено кілька характеристик, які не могли бути оброблені безпосередньо, оскільки вони мали високий відсоток відсутніх значень. Цими характеристиками були вага (97% відсутніх значень), код платника (40%) та медична спеціальність (47%).

Атрибут ваги вважався занадто розрідженим і не був включений у подальший аналіз. Код платника було видалено, оскільки він мав високий відсоток відсутніх значень і не вважався релевантним для результату. Атрибут медичної спеціальності було збережено, додавши значення "відсутнє" для врахування відсутніх значень.

Попередній набір даних містив множинні стаціонарні візити для деяких пацієнтів, і спостереження не можна було вважати статистично незалежними, що є припущенням моделі логістичної регресії. Таким чином, використовувалося лише одне звернення на пацієнта; зокрема, розглядалося лише перше звернення для кожного пацієнта як первинна госпіталізація та визначалося, чи були вони повторно госпіталізовані протягом 30 днів.

Додатково було видалено всі звернення, що закінчилися випискою до хоспісу або смертю пацієнта, щоб уникнути упередженості в аналізі. Після виконання вищеописаних операцій залишилося 69,984 звернень, які склали остаточний набір даних для аналізу.

Змінні, обрані для контролю демографічних характеристик пацієнтів та тяжкості захворювання, включали стать, вік, расу, джерело госпіталізації, характер виписки, первинний діагноз, медичну спеціальність лікаря, що госпіталізував, та час, проведений у лікарні.

Підсумовуючи, набір даних "Diabetes 130-US Hospitals for Years 1999-2008" складається з госпіталізацій тривалістю від одного до 14 днів, які не закінчилися смертю пацієнта або випискою до хоспісу. Кожне звернення відповідає унікальному пацієнту з діагнозом діабету, хоча первинний діагноз може відрізнятися. Під час кожного з аналізованих звернень призначалися лабораторні дослідження та вводилися медикаменти.

4 ОПИС ЕКСПЕРЕМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

4.1 Проведення експериментальних досліджень

Система прогнозування ризику повторної госпіталізації діабетичних пацієнтів побудована за модульною архітектурою, що забезпечує можливість легкого додавання нових алгоритмів та порівняння їх ефективності. Основна логіка системи реалізована у головному модулі, який координує роботу трьох спеціалізованих моделей машинного навчання.

Центральна частина системи включає критично важливий етап підготовки даних:

```
    drop_cols = ['encounter_id', 'patient_nbr', 'weight', 'payer_code',
'medical_specialty']
    df = df.drop(columns=[c for c in drop_cols if c in df.columns])
    df['readmitted_flag'] = np.where(df['readmitted']=='<30', 1, 0)
```

Видалення ознак encounter_id та patient_nbr обумовлено тим, що ці ідентифікатори не несуть предиктивної інформації для моделі. Ознака weight виключена через критично високий відсоток відсутніх значень (97%), що може призвести до значного зменшення розміру навчального набору. Атрибути payer_code та medical_specialty також мають високий відсоток пропущених значень (52% та 53% відповідно) та не вважаються ключовими для медичного прогнозування.

Бінарна цільова змінна readmitted_flag створюється на основі оригінального атрибута readmitted, де значення "<30" (повторна госпіталізація протягом 30 днів) кодується як 1, а всі інші випадки як 0.

```
pythonX_enc = pd.get_dummies(X, drop_first=True)
num_cols = X_enc.select_dtypes(include=[np.number]).columns
scaler = StandardScaler()
X_enc[num_cols] = scaler.fit_transform(X_enc[num_cols])
```

One-hot кодування застосовується для перетворення категоріальних змінних у числовий формат. Параметр drop_first=True використовується для уникнення

мультиколінеарності через dummy variable trap. StandardScaler застосовується виключно до числових ознак для приведення їх до стандартизованого масштабу (середнє = 0, стандартне відхилення = 1), що критично важливо для логістичної регресії та покращує збіжність алгоритмів оптимізації.

```
pythonX_train, X_test, y_train, y_test = train_test_split(
    X_enc, y, test_size=0.2, random_state=42, stratify=y)
```

Співвідношення 80/20 для тренувального та тестового наборів є стандартною практикою в машинному навчанні, що забезпечує достатню кількість даних для навчання при збереженні репрезентативної тестової вибірки. Параметр stratify=y гарантує збереження пропорції класів у обох вибірках, що критично важливо для незбалансованих наборів даних, де клас "повторно госпіталізованих" становить лише близько 11% від загальної кількості спостережень..

4.2 Модель логістичної регресії

Логістична регресія представляє собою лінійний класифікатор, що використовує логістичну (сигмоїдну) функцію для моделювання ймовірності належності до класу:

```
def train_lr(X_train, X_test, y_train, y_test):
    model = LogisticRegression(max_iter=1000, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:, 1]
```

Параметр max_iter=1000 встановлений для забезпечення достатньої кількості ітерацій для збіжності алгоритму оптимізації L-BFGS (за замовчуванням у scikit-learn). Цей вибір обумовлений розміром навчального набору (понад 55,000 спостережень після обробки) та великою кількістю ознак після one-hot кодування. Недостатня кількість ітерацій може призвести до передчасного завершення навчання до досягнення оптимуму.

```
auc = roc_auc_score(y_test, y_prob)
fpr, tpr, _ = roc_curve(y_test, y_prob)
```

Модель повертає як дискретні прогнози (y_pred), так і ймовірності (y_prob), що дозволяє обчислювати як класичні метрики класифікації, так і ROC-аналіз. ROC AUC особливо важливий для медичних додатків, оскільки дозволяє оцінити якість ранжування пацієнтів за ризиком незалежно від конкретного порогу класифікації.

Логістична регресія вибрана як базовий алгоритм через її інтерпретовність - коефіцієнти моделі безпосередньо показують вплив кожної ознаки на логарифм відношення шансів. Це критично важливо в медичних застосуваннях, де необхідно розуміти, які фактори впливають на прогноз.

4.3 Модель дерева рішень

Дерево рішень використовує алгоритм CART (Classification and Regression Trees) з критерієм Джині за замовчуванням. Відсутність явного обмеження глибини (max_depth=None) дозволяє алгоритму будувати дерево до повного розділення вузлів або досягнення мінімальної кількості зразків у листі (min_samples_leaf=1 за замовчуванням).

```
def train_dt(X_train, X_test, y_train, y_test):
    model = DecisionTreeClassifier(random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_prob = model.predict_proba(X_test)[:, 1]
```

Алгоритм рекурсивно розділяє простір ознак, обираючи на кожному кроці ознаку та поріг розділення, що максимально зменшують нечистоту Джині відповідно до формули (4.1):

$$\mathrm{Gini}(S) = 1 - \Sigma(p_i^2) \tag{4.1}$$

де Gini(S) – індекс нечистоти Джині для вузла S;

p_i – частка зразків класу;

i у вузлі S.

Цей підхід дозволяє дереву автоматично виявляти нелінійні залежності та взаємодії між ознаками без попередньої специфікації.

Дерево рішень природним чином обробляє як числові, так і категоріальні ознаки, не потребуючи попередньої нормалізації. Модель генерує ймовірності класифікації на основі розподілу класів у листових вузлах, що робить можливим ROC-аналіз.

4.4 Модель XGBoost

XGBoost реалізує алгоритм екстремального градієнтного бустингу, що будує ансамбль слабких учнів (зазвичай дерев рішень) послідовно, де кожне наступне дерево намагається виправити помилки попередніх.

```
def train_xgb(X_train, X_test, y_train, y_test):
          X_tr = X_train.values
          X_te = X_test.values

          model = XGBClassifier(use_label_encoder=False,
eval_metric='logloss', random_state=42)
          model.fit(X_tr, y_train)
```

Конверсія DataFrame у numpy arrays (X_train.values) здійснюється для уникнення попереджень XGBoost щодо назв стовпців, які можуть містити спеціальні символи після one-hot кодування. Це технічна особливість, що не впливає на якість моделі.

Параметр use_label_encoder=False використовується для уникнення deprecated warning у новіших версіях XGBoost. eval_metric='logloss' явно специфікує метрику для оптимізації, що відповідає логарифмічній функції втрат для бінарної класифікації: (4.2):

$$\text{LogLoss} = -1/N * \Sigma[yi*\log(pi) + (1-yi)*\log(1-pi)] \quad (4.2)$$

де LogLoss – логарифмічна функція втрат;

N – кількість спостережень у вибірці;

yi – фактичне значення цільової змінної для і-го спостереження (0 або 1);

pi – прогнозована ймовірність належності до позитивного класу для і-го спостереження.

Ансамблевий характер алгоритму зазвичай забезпечує вищу предиктивну точність порівняно з окремими моделями, особливо на складних наборах даних з нелінійними залежностями.

4.4 Візуалізація та аналіз результатів

ROC-криві дозволяють візуально порівняти здатність моделей розрізняти класи при різних порогах класифікації. Діагональна лінія представляє випадкове прогнозування (AUC = 0.5), а ідеальний класифікатор мав би AUC = 1.0.

```
plt.figure(figsize=(8,6))
     plt.plot(fpr_lr,       tpr_lr,       label=f'Logistic       Regression
(AUC={auc_lr:.2f})')
     plt.plot(fpr_dt, tpr_dt, label=f'Decision Tree (AUC={auc_dt:.2f})')
     plt.plot(fpr_xgb, tpr_xgb, label=f'XGBoost (AUC={auc_xgb:.2f})')
     plt.plot([0,1],[0,1],'k--')
```

ROC-криві дозволяють візуально порівняти здатність моделей розрізняти класи при різних порогах класифікації. Діагональна лінія представляє випадкове прогнозування (AUC = 0.5), а ідеальний класифікатор мав би AUC = 1.0 (див.рис.4.1. та табл. 4.1).

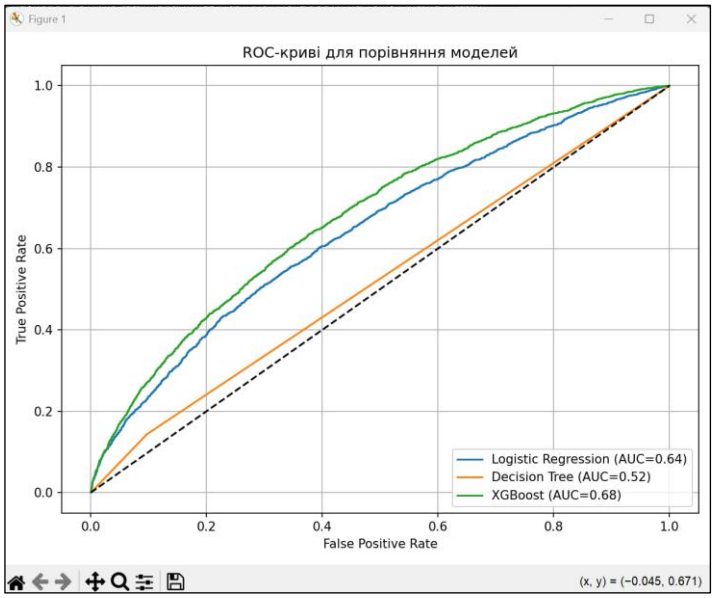

Рисунок 4.1 – Результати порівняння моделей (рисунок створено самостійно)

Таблиця 4.1. Порівняльна характеристика методів (таблицю створено самостійно)

Модель	Accuracy	Precision (клас 1)	Recall (клас 1)	F1-score (клас 1)	ROC AUC	Precision (клас 0)	Recall (клас 0)
Логістична регресія	0.89	0.50	0.02	0.04	0.6446	0.89	1.00
Дерево рішень	0.82	0.16	0.14	0.15	0.5232	0.89	0.90
XGBoost	0.89	0.51	0.02	0.04	0.6793	0.89	1.00

Accuracy (Загальна точність): логістична регресія та XGBoost демонструють однаково високу загальну точність (89%), що значно перевищує результат дерева рішень (82%). Однак ця метрика може бути оманливою для незбалансованих наборів даних, оскільки високе значення може досягатися за рахунок правильної класифікації переважаючого класу (неповторно госпіталізовані пацієнти).

Precision для позитивного класу (повторна госпіталізація): XGBoost показує найвищу точність (0.51), що означає, що 51% пацієнтів, яких модель прогнозує як "високий ризик повторної госпіталізації", дійсно госпіталізуються повторно. Логістична регресія демонструє схожий результат (0.50), тоді як дерево рішень значно поступається (0.16). Висока точність критично важлива для мінімізації хибних тривог у медичному контексті.

Recall для позитивного класу (чутливість): всі три моделі демонструють критично низьку чутливість: логістична регресія і XGBoost виявляють лише 2% пацієнтів, які дійсно будуть повторно госпіталізовані, дерево рішень - 14%. Це означає, що моделі не здатні ефективно ідентифікувати пацієнтів високого ризику, що є серйозною проблемою для клінічного застосування.

F1-score для позитивного класу: низькі значення F1-score (0.04 для логістичної регресії та XGBoost, 0.15 для дерева рішень) відображають дисбаланс між точністю та повнотою. Дерево рішень показує найкращий баланс, хоча абсолютні значення залишаються неприйнятно низькими.

ROC AUC (Площа під ROC-кривою): XGBoost демонструє найвищий ROC AUC (0.6793), що вказує на найкращу здатність ранжувати пацієнтів за ризиком. Логістична регресія показує середній результат (0.6446), тоді як дерево рішень

практично не відрізняється від випадкового прогнозування (0.5232). ROC AUC є найбільш релевантною метрикою для медичних застосувань, оскільки дозволяє оцінити якість ранжування незалежно від порогу класифікації.

Специфічність (Recall для класу 0): логістична регресія та XGBoost демонструють ідеальну специфічність (1.00), правильно ідентифікуючи всіх пацієнтів з низьким ризиком. Дерево рішень показує дещо нижчий результат (0.90), але все ще високий.

XGBoost демонструє найкращу загальну ефективність за ROC AUC та може використовуватися для ранжування пацієнтів за ризиком.Дерево рішень, незважаючи на нижчі загальні метрики, показує найкращий баланс між точністю та повнотою для позитивного класу.Логістична регресія забезпечує найкращу інтерпретовність результатів, що важливо для клінічного прийняття рішень

Поточні результати вказують на необхідність подальшої оптимізації для досягнення клінічно прийнятної ефективності у виявленні пацієнтів високого ризику повторної госпіталізації.

ВИСНОВКИ

У ході виконання кваліфікаційної роботи було проведено комплексний аналіз проблемної області застосування моделей машинного навчання для прогнозування ризику розвитку цукрового діабету та раннього виявлення медичних ризиків. На основі описаних експериментів з трьома базовими алгоритмічними підходами – логістичною регресією, деревом рішень та XGBoost – отримано важливі наукові та практичні результати.

Проведене дослідження підтвердило гіпотезу про різну ефективність алгоритмів машинного навчання при розв'язанні задач медичної діагностики. XGBoost продемонстрував найвищу прогностичну здатність з ROC AUC 0.6793 та найкращою точністю класифікації позитивного класу (0.51), що робить його оптимальним для ранжування пацієнтів за ризиком. Логістична регресія показала порівнянну загальну точність (89%) та забезпечила найкращу інтерпретовність результатів, критично важливу для клінічного прийняття рішень. Дерево рішень, незважаючи на нижчі загальні метрики, виявило найкращий баланс між точністю та повнотою для виявлення пацієнтів високого ризику.

Критичною проблемою всіх досліджених моделей є низька чутливість (2-14%) до позитивного класу, що вказує на обмежену здатність ефективно ідентифікувати пацієнтів високого ризику. Це пов'язано з характерним для медичних даних дисбалансом класів. Отримані результати демонструють необхідність подальшої оптимізації моделей, включаючи техніки обробки незбалансованих даних, feature engineering та ансамблеві методи.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Babyak, Michael A. “What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models.” *Psychosomatic Medicine* 66, no. 3 (2004): 411–421.

2. Bellman, Richard E. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press, 1961.

3. Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1984.

4. Chen, Tianqi, and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, 785–794. New York: ACM, 2016. https://doi.org/10.1145/2939672.2939785.

5. Choi, Edward, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. “Using Recurrent Neural Network Models for Early Detection of Heart Failure Onset.” *Journal of the American Medical Informatics Association* 24, no. 2 (2016): 361–370. https://doi.org/10.1093/jamia/ocw112.

6. Cruz, J. A., and D. S. Wishart. “Applications of Machine Learning in Cancer Prediction and Prognosis.” *Cancer Informatics* 2 (2006): 59–77. https://doi.org/10.1177/117693510600200030.

7. Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher. “International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease.” *American Journal of Cardiology* 64, no. 5 (1989): 304–310. https://doi.org/10.1016/0002-9149(89)90524-9.

8. Deo, Rahul C. “Machine Learning in Medicine.” *Circulation* 132, no. 20 (2015): 1920–1930.

9. Dormann, Carsten F., et al. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” *Ecography* 36, no. 1 (2013): 27–46.

10. Efron, Bradley, and Robert Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

11. Harrell Jr, Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham: Springer, 2015.

12. Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.

13. International Diabetes Federation. *IDF Diabetes Atlas*, 10th ed. Brussels: IDF, 2021. https://diabetesatlas.org/.

14. International Diabetes Federation. "Diabetes Now Affects One in 10 Adults Worldwide." *IDF News*, November 2, 2021. https://idf.org/news/240:diabetes-now-affects-1-in-10-adults-worldwide.html.

15. Japkowicz, Nathalie, and Shaju Stephen. "The Class Imbalance Problem: A Systematic Study." *Intelligent Data Analysis* 6, no. 5 (2002): 429–449.

16. Kleinbaum, David G., and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. 3rd ed. New York: Springer, 2010.

17. Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. 5th ed. New York: McGraw-Hill, 2005.

18. Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 4768–4777, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

19. Miotto, Riccardo, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. "Deep Learning for Healthcare: Review, Opportunities and Challenges." *Briefings in Bioinformatics* 19, no. 6 (2018): 1236–1246. https://doi.org/10.1093/bib/bbx044.

20. Moons, Karel GM, et al. "Risk Prediction Models: I. Development, Internal Validation, and Assessing the Incremental Value of a New (Bio)marker." *Heart* 98, no. 9 (2012): 683–690.

21. Nori, V. S., B. Hane, M. G. Sun, A. H. Gillies, R. J. Rybicki, and K. R. Natarajan. "Deep Neural Network Models for Identifying Incident Dementia Using

Claims and EHR Datasets." *PLoS ONE* 15, no. 5 (2020): e0233973. https://doi.org/10.1371/journal.pone.0233973.

22. Quinlan, J. R. "Induction of Decision Trees." *Machine Learning* 1, no. 1 (1986): 81–106. https://doi.org/10.1007/BF00116251.

23. Quinlan, J. R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

24. Rajkomar, Alvin, Eyal Oren, Kai Chen, et al. "Scalable and Accurate Deep Learning with Electronic Health Records." *npj Digital Medicine* 1 (2018): 18. https://doi.org/10.1038/s41746-018-0029.

25. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You? Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144. New York: ACM, 2016. https://doi.org/10.1145/2939672.2939778.

26. Rokach, Lior, and Oded Maimon. "Top-Down Induction of Decision Trees Classifiers—A Survey." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35, no. 4 (2005): 476–487. https://doi.org/10.1109/TSMCC.2004.843247.

27. Sherer, Paul M. "Ukraine Receives Seven-Week Supply of Long-Acting Insulin from Direct Relief." *Direct Relief*, 2022. https://www.directrelief.org.

28. Steyerberg, Ewout W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. New York: Springer, 2019.

29. Taylor, R. A., D. Pare, S. A. Venkatesh, L. Mowafi, E. Melnick, M. Fleischman, and H. P. Hall. "Prediction of In-Hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data–Driven, Machine Learning Approach." *Academic Emergency Medicine* 23, no. 3 (2016): 269–278. https://doi.org/10.1111/acem.12876.

30. Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. New York: Springer, 2012.

31. World Health Organization. “Treating Diabetes: WHO Delivers Insulin to Hospitals in Ukraine.” *WHO Newsroom*, September 1, 2022. https://www.who.int/news-room.

32. Yang, G., and Jun Feng. “Federated Learning for Healthcare: Systematic Review and Future Directions.” *IEEE Transactions on Artificial Intelligence* 3, no. 1 (2022): 1–12. https://doi.org/10.1109/TAI.2022.3141824.

33. “At Least 1.2 Million Ukrainians Have Been Diagnosed with Diabetes – Health Minister.” *Bukvy*, November 14, 2024. https://bykvu.com.

34. GitHub – KatePotomkina / disease_prediction . *GitHub*. URL: https://github.com/KatePotomkina/disease_prediction/tree/master (дата звернення: 15.06.2025).

35. Smelyakov, K., O. Klochko, and Z. Dudar. 2023. “Building Quantile Regression Models for Predicting Traffic Flow.” Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems.

36. Smelyakov, K., P. Dmitry, M. Vitalii, and C. Anastasiya. 2018. “Investigation of Network Infrastructure Control Parameters for Effective Intellectual Analysis.” In Proceedings of the 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 983–986. https://doi.org/10.1109/TCSET.2018.8336359.

37. Smelyakov, K., A. Chupryna, D. Sandrkin, and M. Kolisnyk. 2020. “Search by Image Engine for Big Data Warehouse.” In 2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, 1–4. https://doi.org/10.1109/eStream50540.2020.9108782.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ ЗА НАУКОВИМИ НАПРЯМАМИ КЕРІВНИКА ТА НАУКОВЦІВ КАФЕДРИ ПРОГРАМНОЇ ІНЖЕНЕРІЇ

35. Smelyakov, K., O. Klochko, and Z. Dudar. 2023. “Building Quantile Regression Models for Predicting Traffic Flow.” *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Systems*.

36. Smelyakov, K., P. Dmitry, M. Vitalii, and C. Anastasiya. 2018. “Investigation of Network Infrastructure Control Parameters for Effective Intellectual Analysis.” In *Proceedings of the 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, Lviv-Slavske, Ukraine, 983–986. https://doi.org/10.1109/TCSET.2018.8336359.

37. Smelyakov, K., A. Chupryna, D. Sandrkin, and M. Kolisnyk. 2020. “Search by Image Engine for Big Data Warehouse.” In *2020 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania, 1–4. https://doi.org/10.1109/eStream50540.2020.9108782.